

Композиции решающих функций

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».
Лекция 6.

Задача построения решающей функции

Пусть X – пространство значений переменных,
используемых для прогноза,
 $Y = \{-1, 1\}$ – пространство значений прогнозируемых
переменных,

C – множество всех вероятностных мер на заданной
 σ -алгебре подмножеств множества $D = X \times Y$.

Решающей функцией (алгоритмом классификации)
называется соответствие $\lambda: X \rightarrow Y$.

Отображение $Q: D^N \rightarrow \Lambda$ множества выборок во множество
решающих функций называется методом (алгоритмом)
построения решающих функций.

Композиции классификаторов

Обобщение решающей функции: $\lambda: X \rightarrow [0, +\infty)$ — вводится пространство оценок.

Пусть имеются T решающих функций $\lambda_1(x), \dots, \lambda_T(x)$.

Композиция есть решение в виде

$$\lambda(x) = C(\lambda_1(x), \dots, \lambda_T(x)),$$

где $C(\cdot, \dots, \cdot)$ — монотонна по всем аргументам.

Функции $\lambda_t(x)$ принимают значения из пространства оценок, значения функции $\lambda(x)$ — из множества Y .

Линейные композиции

Линейная композиция

$$\lambda(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t \lambda_t(x) \right), \quad \alpha_t \geq 0.$$

Методы построения композиций

- Бустинг (AdaBoost, градиентный бустинг)
- Бэггинг и метод случайных подпространств
- Комитетные методы
- Голосование (простое, взвешенное, по старшинству)

Смеси алгоритмов – если α_t зависит от x , идея областей компетентности.

Алгоритм AdaBoost

В методе AdaBoost решение строится в виде композиции

$$\lambda(x) = \text{sign}(\beta(x)), \quad \beta(x) = \sum_{t=1}^T \alpha_t \lambda_t(x),$$

где базовые классификаторы $\lambda_t(x)$ и их веса α_t находятся следующим образом.

Первый базовый классификатор строится базовым методом на основе исходной выборки, объектам которой приписаны начальные веса $w^1 = (w_1^1, \dots, w_N^1)$.

Заметим, что мы будем задавать начальные веса объектам в соответствии с выбранным распределением, но в стандартном варианте метода начальные веса выбираются одинаковыми, т.е. $w_i^1 = \frac{1}{N}$.

Пересчёт весов

Вес построенного базового классификатора в композиции определяется по формуле

$$\alpha_t = \frac{1}{2} \ln \frac{\widetilde{M}^+(V, w^t, \lambda_t)}{\widetilde{M}^-(V, w^t, \lambda_t)},$$

где

$$\widetilde{M}^+(V, w, \lambda) = \sum_{i=1}^N w_i \cdot I(y^i = \lambda(x^i)),$$

$$\widetilde{M}^-(V, w, \lambda) = \sum_{i=1}^N w_i \cdot I(y^i = -\lambda(x^i)).$$

Итерационный процесс

Следующие базовые классификаторы строятся тем же базовым методом по выборке, веса объектов в которой вычисляются по формулам

$$w_i^{t+1} = \frac{\bar{w}_i^{t+1}}{\sum_{i=1}^N \bar{w}_i^{t+1}}, \quad \bar{w}_i^{t+1} = w_i^t \cdot e^{-\alpha_t y^i \lambda_t(x^i)}.$$

Веса правильно классифицированных объектов умножаются на $e^{-\alpha_t}$, а веса неправильно классифицированных объектов умножаются на e^{α_t} .

Случай независимых переменных

Из формулы Байеса можем записать

$$g(x) = P(y = 1 \mid x) = \frac{P(dx, y = 1)}{P(dx, y = 1) + P(dx, y = -1)}$$

$$g(x) = \frac{1}{1 + \frac{1-p}{p} \cdot \frac{P(dx|y=-1)}{P(dx|y=1)}}.$$

Пусть условные распределения всех переменных X_j при условии обоих классов независимы, т.е.

$$P(dx \mid y) = \prod_{j=1}^n P(dx_j \mid y).$$

Сведение к логистической функции

Подставив это произведение в предыдущее выражение, после преобразований имеем

$$\frac{p}{1-p} \cdot \left(\frac{1}{g(x)} - 1 \right) = \prod_{j=1}^n \frac{p}{1-p} \cdot \left(\frac{1}{g_j(x_j)} - 1 \right),$$

где $g_j(x_j) = P(y = 1 \mid x_j) = \frac{P(dx_j, y=1)}{P(dx_j)}$.

Логарифмируем последнее выражение и получаем

$$\sigma^{-1}(g(x)) = (n-1)(\ln p - \ln(1-p)) + \sum_{j=1}^n \sigma^{-1}(g_j(x_j)),$$

где $\sigma^{-1}(\cdot)$ — функция, обратная сигмоиду $\sigma(z) = \frac{1}{1+e^{-z}}$.

Обобщённый наивный байесовский классификатор

Заметим, что полученное выражение имеет вид логистической регрессии, а именно

$$g(x) = \sigma \left(u_0 + \sum_{j=1}^n u_j \sigma^{-1}(g_j(x_j)) \right),$$

при $u_0 = (n - 1)(\ln p - \ln(1 - p))$, $u_j = 1$.

Обычно логистическую кривую получают, исходя из предположений о виде распределения, однако сейчас мы предположили независимость переменных, но не ограничивали вид распределений.

Использование модели

Данное выражение справедливо не только при независимых переменных, а в несколько более общем случае, поскольку из предыдущего соотношения независимость переменных не следует.

Ещё более расширить область применимости можно, если считать веса свободными параметрами.

Дальнейшее обобщение возможно, если допустить произвольные оценочные функции

$$g(x) = \sigma \left(u_0 + \sum_{j=1}^n u_j s(x_j) \right).$$

Оценивание условной вероятности

Условную вероятность $g(x) = P(y = 1 | x)$ представим как находящиеся в точке x два объекта: класса 1 с весом $w_0 g(x)$ и класса -1 с весом $w_0(1 - g(x))$.

В результате выполнения бустинга вес первого объекта станет равным

$$w^{+1}(x) = w_0 g(x) \cdot A e^{-\beta(x)},$$

где константа A есть произведение всех нормировочных множителей.

Конечный вес второго объекта есть

$$w^{-1}(x) = w_0(1 - g(x)) \cdot A e^{\beta(x)}.$$

Если приравнять веса объектов, то получим

$$g(x) = \frac{1}{1 + e^{-2\beta(x)}}.$$

Бустинг на пороговых классификаторах

Бустинг на пороговых классификаторах («пнях») является разновидностью обобщённого наивного байесовского классификатора.

Действительно, каждая $\lambda_t(x)$ в композиции

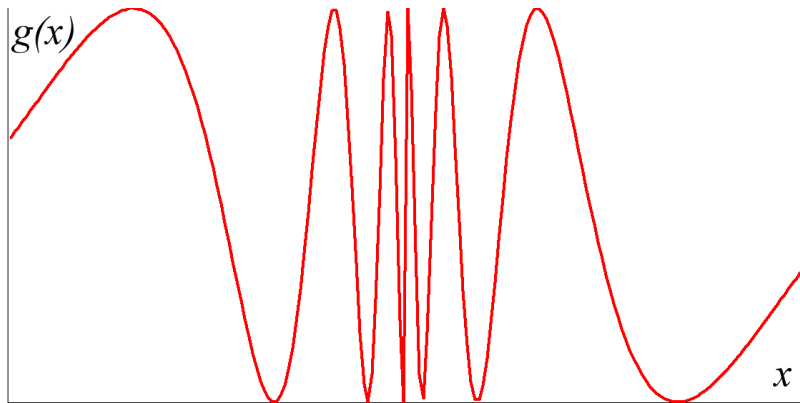
$$\beta(x) = \sum_{t=1}^T \alpha_t \lambda_t(x)$$

зависит только от одной переменной X_{i_t} , поэтому после группировки слагаемых выражение можно привести к виду

$$2\beta(x) = \sum_{i=1}^n u_i s(x).$$

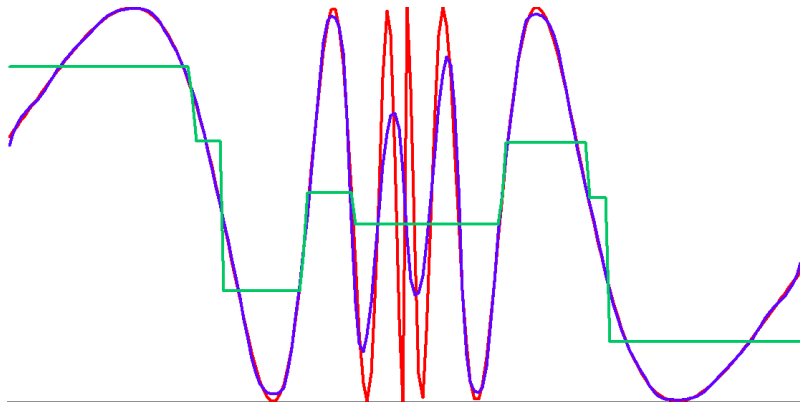
Подставив в выражение для $g(x)$, получим искомый вид.

Модельный пример



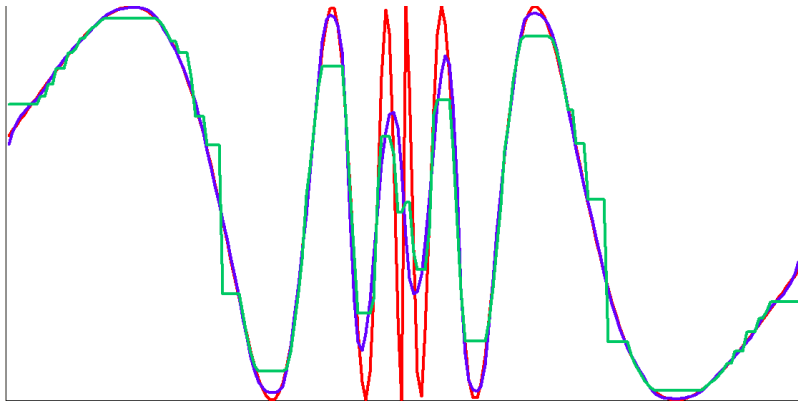
Функция условной вероятности.

Аппроксимация сплайном



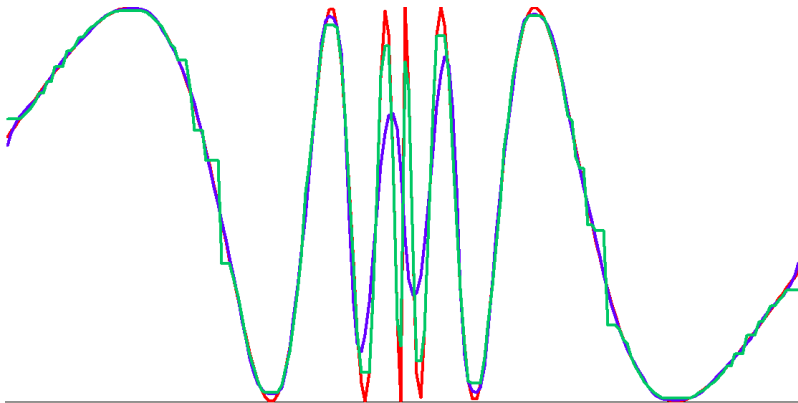
Кубический сплайн на 20 интервалов.
AdaBoost 10 итераций.

Boosting



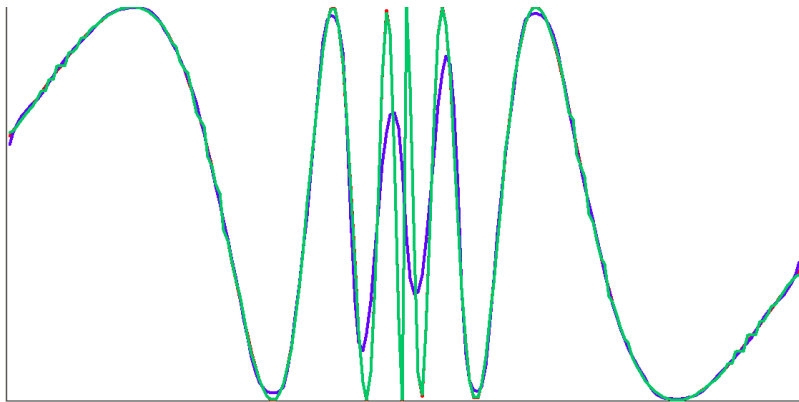
AdaBoost 100 итераций.

Boosting



AdaBoost 1000 итераций.

Boosting



AdaBoost 10000 итераций.

Понятие отступа

Отступ есть

$$\theta = \frac{y\beta(x)}{\varkappa}, \quad \varkappa = \sum_{t=1}^T \alpha_t.$$

Из-за нормировки в виде \varkappa сложность композиции влияет на оценку риска.

Выводы

- Важнейшей причиной эффективности бустинга является использование эффекта независимости (переменных, подпространств, моделей).
- Бустинг на пороговых классификаторах является разновидностью непараметрической логистической регрессии, также его можно считать разновидностью (существенно обобщённого) наивного байесовского классификатора.
- Бустинг реализует «удачный» вариант непараметрической аппроксимации условной вероятности.